Original Paper

Beyond Area Under the Receiver Operating Characteristic Curve: Evaluating Predictive Performance Metrics Under Class Imbalance in Real-World Clinical Data

Vanessa das Graças José Ventura[1], MSc, MD; Claudio Moisés Valiense de Andrade[2], BSc, MSc; Jussara Marques de Almeida[2], MSc, PhD; Bruno Porto Pessoa[3], MSc, MD; Carísi Anne Polanczyk[4,5,6], MSc, MD, PhD; Guilherme Fonseca do Nascimento[2], BSc; Eric Boersma[7], MSc, PhD; Heloisa Reniers Vianna[8], MSc, MD; Katia de Paula Farah[1], MSc, MD, PhD; Leonardo Chaves Dutra da Rocha[2], MSc, PhD; Marcos André Gonçalves[2], MSc, PhD; Milena Soriano Marcolino[5,9,10], MSc, MD, PhD

[1]Medical School and University Hospital, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
[2]Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
[3]Hospital Júlia Kubitschek, Belo Horizonte, Brazil
[4]Department of Medicine Internal, Medical School, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil
[5]Institute for Health Assessment and Translation for Chronic and Neglected Diseases of High RElevance (IATS-CARE), Belo Horizonte, Brazil
[6]Hospital Moinhos de Vento, Porto Alegre, Brazil
[7]Erasmus University Medical Center, Rotterdam, The Netherlands
[8]Hospital Universitário Ciências Médicas, Belo Horizonte, Brazil
[9]Department of Internal Medicine, Medical School, Medical School and University Hospital, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
[10]Telehealth Center, University Hospital, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Corresponding Author:

Vanessa das Graças José Ventura, MSc, MD
Medical School and University Hospital
Universidade Federal de Minas Gerais
Avenida Alfredo Balena, 110
Belo Horizonte 30130-100
Brazil
Phone: 55 31991314221
Email: nessachesed@yahoo.com.br

Abstract

Background: Predictive models increasingly support clinical decision-making, although imbalanced outcome distributions are common in health care datasets and can distort performance evaluation. The area under the receiver operating characteristic curve (AUROC) remains the most frequently reported metric, despite its limited ability to reflect clinically meaningful performance under class imbalance.

Objective: This study aimed to examine the influences of metric selection on the clinical interpretation of predictive models in imbalanced real-world health care data.

Methods: This was a retrospective cohort study, including 17,018 hospitalized patients with COVID-19. Two predictive models using extreme gradient boosting (XGBoost) were developed to predict kidney replacement therapy (KRT) and mortality. Model performance was assessed using AUROC, macro-F_1-score, class-specific precision and recall, calibration (curve, slope, and intercept), decision curve analysis, and learning curves. Standard rebalancing strategies were applied exclusively to the training data to evaluate their impact on performance.

Results: KRT occurred in 9.5%, and mortality in 18.0%. Although AUROC values were high (0.928 for KRT and 0.945 for mortality), performance in the minority class was substantially lower. For KRT, precision was 0.539 and recall 0.372; for mortality, precision was 0.725 and recall 0.718. Rebalancing strategies were associated with higher recall for the minority class, but this gain was accompanied by a reduction in precision, with minimal impact on AUROC values. As a result, AUROC remained high despite clinically relevant changes in error distribution between false positives and false negatives. The learning curves show a plateau-like shape, with stable validation performance across all training set sizes for both outcomes.

Conclusions: AUROC alone is insufficient to evaluate prediction models in imbalanced health care scenarios, even with rebalancing. Routine reporting of class-aware metrics, alongside learning curve analysis, is essential to support robust and clinically meaningful evaluation of predictive models, rather than their direct translation into practice.

JMIR Form Res 2026;10:e86379; doi: 10.2196/86379

Keywords: predictive model; artificial intelligence; learning curve; area under the receiver operating characteristic curve; F_1-score; performance metrics

Introduction

Clinical prediction models are increasingly used in health care to support diagnostic, prognostic, and therapeutic decisions [1,2]. Their adoption has expanded with advances in machine learning (ML) and access to large-scale electronic health data, enabling the development of models with the potential to improve risk stratification and personalized care [3]. However, the evaluation of these models frequently relies on metrics that may not reflect real-world clinical usefulness, especially when outcome distributions are imbalanced [4].

The area under the receiver operating characteristic curve (AUROC) is the most commonly reported metric in clinical prediction research [5,6]. Although AUROC is intuitive and threshold-independent, it often overestimates performance in datasets in which one class (eg, survival or absence of disease) predominates, a common characteristic in clinical datasets [7-9]. In such settings, AUROC may suggest high discriminative ability while concealing poor sensitivity for minority-class outcomes, such as death or the need for critical interventions [7-9].

For example, risk prediction scores have long been recommended in clinical practice to guide preventive strategies, particularly in cardiovascular disease. The Framingham score, once widely used for predicting 10-year cardiovascular outcomes, illustrates the risk of relying on AUROC-based metrics [10]. Although it showed acceptable discrimination (C-statistic: 0.763 for men and 0.793 for women) [10,11], the dataset exhibited class imbalance (10.08% women and 18.09% men with outcomes) [10], and the score likely performed better for healthy individuals while failing to identify many at-risk patients early, potentially missing opportunities for interventions that could have improved outcomes [12-14].

For this reason, calibration measures are essential complements to discrimination, since high AUROC values do not necessarily guarantee reliable probability estimates or clinical applicability [6,15,16]. More recent cardiovascular risk prediction models, such as SCORE2, incorporated improved calibration across European populations, but their performance reporting still relies heavily on AUROC [17].

This illustrates a broader issue: even when calibration is addressed, discrimination metrics alone can mask poor identification of minority outcomes, underscoring the need for comprehensive evaluation strategies. Additionally, although the Hosmer-Lemeshow test is a commonly used goodness-of-fit test for logistic regression models, it is less suitable for ML models due to its sensitivity to sample size and arbitrary grouping of predicted probabilities [18].

While these limitations are widely recognized in the ML literature, clinical studies continue to prioritize AUROC in model reporting [19-21]. Most discussions on metric limitations remain either theoretical or based on synthetic datasets [22-25]. As a result, model outputs often lack interpretability and applicability for health professionals, limiting their practical relevance for clinical implementation [26,27]. There are few applied studies using large real-world clinical datasets that demonstrate, in concrete terms, how metric selection affects the identification of high-risk patients and subsequent clinical decisions [4,22,23,27].

Recently, Carriero et al [28] reported the challenges posed by imbalanced datasets in predictive modeling, showing that common strategies to deal with class-imbalance issues, such as oversampling and undersampling, may compromise calibration, leading to overestimated risk predictions and systematic bias [28]. These findings highlight the need for evaluation strategies that move beyond AUROC and artificial rebalancing, offering instead a comprehensive assessment of model performance that prioritizes clinical reliability and patient safety.

This study addresses this gap by applying a structured evaluation of predictive model performance in a real-world clinical setting, using a large, multicenter dataset of hospitalized patients with COVID-19 in Brazil. As a case study to illustrate the impact of class imbalance on model evaluation, we developed 2 ML models to predict kidney replacement therapy (KRT) and in-hospital mortality, outcomes with different prevalence levels, and assessed them using metrics that capture different aspects of model performance. Rather than proposing new predictive models, this study focuses on how commonly used performance metrics influence the interpretation of model usefulness in real-world, imbalanced clinical settings. Beyond AUROC, we focused on class-specific precision, recall, and macro-F_1-scores, which, although well-established in data science, remain underutilized in clinical contexts. Additionally, we critically examine how metric selection influences clinical interpretation in imbalanced scenarios, making our findings relevant not only to data scientists but also to health care professionals. In doing so, this study helps bridge the gap between methodological rigor and clinical applicability in predictive model evaluation.

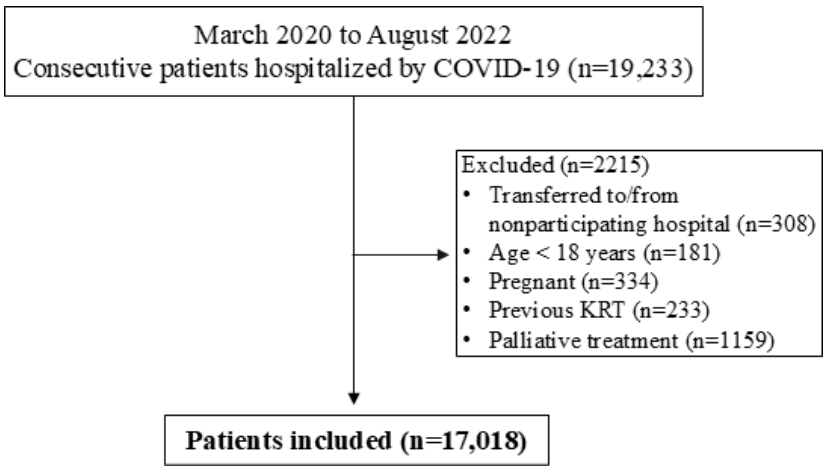
Methods

Study Design

This was a retrospective cohort study. We collected data on consecutive adult patients (aged 18 years and older) with laboratory-confirmed COVID-19 [29], admitted in one of 41 participating hospitals in Brazil from March 2020 to August 2022. Details of the cohort have been published elsewhere [30]. Pregnant women; patients undergoing palliative treatment, or with a history of prior KRT or already in KRT upon hospital presentation; and those who were transferred from or to another hospital were excluded from this particular analysis (Figure 1).

Two predictive models were developed and validated: 1 for KRT and 1 for in-hospital mortality. Both models presented imbalanced class distributions, but in different proportions, and were used as case studies. These outcomes were selected due to their clinical relevance and prognostic implications in hospitalized patients with COVID-19 [31].

Figure 1. Flowchart of the patients included in the study. KRT: kidney replacement therapy.



Data Collection

Sociodemographic, clinical, and laboratory data; medications; interventions; and outcomes were extracted from medical records by trained researchers using the REDCap (Research Electronic Data Capture) electronic platform [32,33], hosted at the Telehealth Center of the University Hospital of the *Universidade Federal de Minas Gerais* [34,35]. An automated data verification algorithm was implemented to ensure data quality, checking for inconsistencies. Any discrepancies were resolved in consultation with the coordinating researchers.

Predictors and Outcome Definition

Candidate predictors were selected based on clinical relevance, prior literature, and data availability (Multimedia Appendices 1 and 2). No automated feature selection was applied, as the study objective was not to maximize predictive performance but to assess how different evaluation metrics behave under identical modeling conditions. The same predictor set was maintained across all experiments to ensure comparability between models and evaluation strategies. KRT was defined as the initiation of dialysis during hospitalization, excluding patients with preexisting chronic dialysis. In-hospital mortality refers to death occurring during hospitalization, as documented in medical records. Patients were classified into binary outcome groups for each end point (KRT vs no KRT; death vs survival).

The Predictive Models

Extreme gradient boosting (XGBoost) was chosen due to its strong performance in structured clinical data, ability to capture nonlinear relationships, native handling of missing values, and favorable calibration properties reported in prior studies [36-38]. Since XGBoost supports missing values natively, no imputation method was used in the primary analysis.

Cross-Validation and Modeling Pipeline

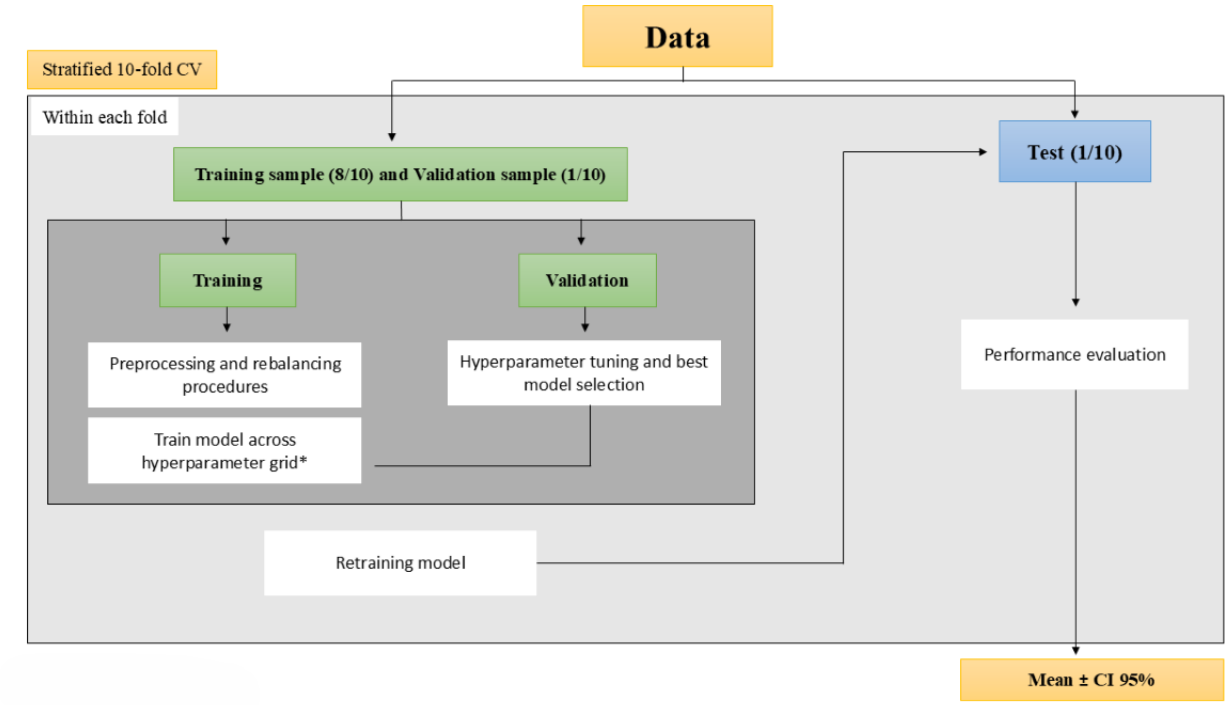
A 10-fold stratified cross-validation strategy was used. In each iteration, 1-fold was held out as the test set, while the remaining 9-folds constituted the training set. Within this training partition, a further split was performed to create a validation subset used exclusively for hyperparameter tuning and model selection. All preprocessing and rebalancing procedures that did not require imputation, which are detailed in the next subsection, were performed strictly within the training data of each fold. The test set remained fully held out and was used only for final performance evaluation, preserving the original data distribution and preventing data leakage.

Three key hyperparameters were systematically explored: (1) booster (gbtree, gblinear, and dart), which defines the base learner used to build the ensemble; (2) eta (learning rate), which controls the step-size shrinkage during boosting to prevent overfitting by making the learning process more conservative; and (3) max_depth, which controls the maximum depth of individual trees, thereby regulating model

complexity. The complete grid of values evaluated for each hyperparameter is reported in Multimedia Appendix 3.

After selecting the optimal hyperparameter configuration, the model was retrained on the full training data (training + validation) following standard practice and prior work [39,40], as well as the default behavior of widely used libraries such as scikit-learn (GridSearchCV with refit=True). The held-out test fold was then used exclusively for final performance evaluation. This process was repeated across all folds, so that each fold served once as the test set, and the reported results correspond to the average performance across the 10 iterations. This strategy ensures robust performance estimation and minimizes data leakage [41,42]. The overview of the analytical pipeline is presented in Figure 2.

Figure 2. Overview of the analytical pipeline applied within each cross-validation iteration. *booster, eta, max_depth. CV: cross-validation.



Handling of Class Imbalance

To assess the impact of data imbalance on model performance, both oversampling and undersampling techniques were applied exclusively on the training data. For each method, default resampling parameters were used, without additional tuning, following the standard implementation of each algorithm. Due to the intrinsic operational characteristics of certain resampling methods, a fully balanced (1:1) class distribution was not always achieved. The final class proportions after resampling are presented in Multimedia Appendix 3.

Oversampling techniques included Random Oversampling [43], Adaptive Synthetic [44], Synthetic Minority Oversampling Technique (SMOTE) [45], BorderlineSMOTE [46], SVMSMOTE [47], and KMeansSMOTE [48], which increase minority-class representation through random duplication or synthetic sample generation [43,49]. Undersampling methods included Random Undersampling [44], Redundancy-Based Undersampling [39], e2sc-us (Effective, Efficient, and Scalable Confidence-Based-UnderSampling) [39], Condensed Nearest Neighbor [50], Near Miss 1 [51], and Near Miss 2 [51], which reduce majority-class instances while attempting to preserve relevant decision boundaries [39].

Because most resampling algorithms cannot handle missing values, the MissForest imputation method was incorporated into the pipeline when required, exclusively within the training data for rebalancing experiments [52]. Importantly, the imputation model was fitted exclusively on the training partition within each cross-validation fold and subsequently applied to the corresponding validation and test sets, thereby preventing information leakage.

For each resampling strategy, the entire modeling pipeline, including imputation (when applicable), resampling, and model training, was re-executed within each cross-validation iteration. Hyperparameter tuning was performed from scratch for each resampled dataset, ensuring that model optimization was specific to each data configuration. Details of the hyperparameter search space are provided in Multimedia Appendix 4. No information from the test set was used at any stage of model development, including imputation, resampling, or hyperparameter tuning.

To enable direct comparison across models, a fixed probability threshold of 0.5 was used for classification. This threshold was selected based on its consistent performance across preliminary analyses. All experiments were conducted using a fixed random seed (random_state=42), applied consistently across all stochastic components of the pipeline, including cross-validation splitting, imputation, resampling procedures, and model initialization, ensuring full reproducibility.

Performance Evaluation

Model performance was assessed using a complementary set of metrics that offer different perspectives to evaluate model performance, capturing global discrimination, class-specific behavior, calibration, and clinical usefulness (Multimedia Appendices 5 and 6). All metrics were computed on the held-out test fold in each iteration and averaged across folds. Specifically, we analyzed global metrics, such as accuracy and AUROC, as well as metrics that are more sensitive to class imbalance, such as macro-F_1 and per-class precision and recall, including precision and recall across both majority and minority classes and the impact of different decision thresholds on the values of these metrics.

We first evaluated global performance using accuracy and AUROC. These metrics summarize overall discrimination across all instances but treat all elements equally, regardless of their class, which inherently biases these metrics toward the majority class in imbalanced datasets [2,4]. Accuracy represents the proportion of correctly classified instances [4], while AUROC quantifies the model's ability to rank positive cases higher than negative ones across all decision thresholds [6,11]. However, what is considered a "correct" prediction depends on the chosen decision threshold for the risk score, as different thresholds influence the balance between sensitivity and specificity [4].

To explicitly capture performance under class imbalance, we additionally reported per-class precision, recall, and F_1-score, as well as macro-F_1, which assigns equal weight to each class regardless of prevalence [50,51,53]. To better characterize performance under class imbalance, we first examined class-specific precision and recall, which directly quantify errors for both minority and majority outcomes. Regarding the positive class, recall (sensitivity) reflects the proportion of true cases correctly identified, whereas precision (positive predictive value) reflects the proportion of predicted positives that are true events. The same logic applies to the negative class, where recall corresponds to specificity and precision to negative predictive value [54,55].

In addition to these class-specific metrics, we evaluated resampling strategies using TPRGap, a bias-oriented measure that quantifies performance disparity between classes as the absolute difference between their true-positive rates. This metric directly captures classifier favoritism toward the majority class, which may persist even when global performance measures remain high [56]. Finally, to summarize the trade-off between precision and recall in a single indicator, we reported the F_1-score and its macroaveraged form, which assigns equal weight to each class and is therefore robust to outcome imbalance [50,51,54,57].

While this perspective is common in the ML literature, it may be less intuitive for health care professionals, who are generally more familiar with metrics such as sensitivity and specificity. By reporting both precision and recall for each class, we provide a more nuanced and clinically interpretable understanding of model performance, especially relevant in the presence of class imbalance [50,51,54,57]. This approach enables assessment not only of how well the model identifies patients at risk but also how confidently it excludes those unlikely to experience the outcome. Therefore, it supports a more comprehensive assessment of predictive usefulness and more informed decision-making in clinical applications.

Primary analyses were conducted using a default probability threshold of 0.5, consistent with standard binary classification practice. To explore clinically relevant trade-offs between missed events and false alarms, we further evaluated precision-recall behavior across varying decision thresholds using precision-recall curves [50,56]. The precision-recall curve was generated by plotting precision against recall at various decision thresholds [54].

Model calibration was assessed using the plot with predicted probability against observed probability, testing intercept equals zero and slope equals 1. In a well-calibrated model, there is agreement between observed and predicted events, allowing the probability to be interpreted as the confidence in the prediction [58,59]. In addition, the global accuracy of the model was assessed using the Brier score. The Brier score ranges from 0 to 1, with lower values indicating better probabilistic accuracy [15].

Clinical usefulness was assessed through decision curve analysis, which quantifies net benefit across a range of decision thresholds compared with "treat-all" and "treat-none" strategies [55,57]. While decision curves assess whether model-guided decisions outperform simple strategies, they do not ensure balanced error distribution or detect bias toward the majority class, reinforcing the need for class-specific performance metrics [55,57]. Finally, the learning curves were used as a graphical representation of how a model's performance evolves as training data are added [60].

Risk-of-Bias Assessment and Reporting

This study adheres to the TRIPOD+AI (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence) standards for transparent reporting (Multimedia Appendix 7) [6]. To ensure methodological rigor, we used the PROBAST+AI (Updated Quality, Risk of Bias, and Applicability Assessment Tool for Prediction Models Using Regression or Artificial Intelligence Methods) to assess risk of bias and applicability (Multimedia Appendix 8). The study was considered to have a low risk of bias in all domains (participants, predictors, outcomes, and analysis). However, the lack of external validation of the model should be considered as a point of attention in the domain of analysis. Applicability concerns were judged to be low across all domains [16].

Ethical Considerations

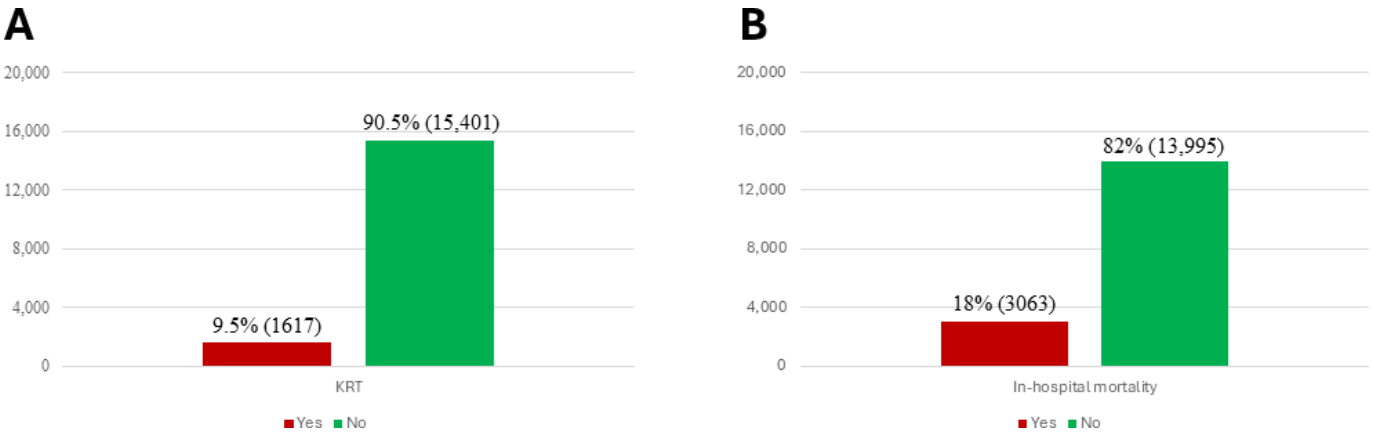
The study was approved by the Brazilian National Research Ethics Committee—Comissão Nacional de Ética em Pesquisa (CAAE 30350820.5.0000.0008) and internal approval of ethics boards from each hospital. Individual informed consent term was waived due to the pandemic situation and analysis of deidentified data, based on chart review only.

Results

Overview

The database included 17,018 patients (median age of 60 years, IQR 37-83; 54.6% were men). The outcome distributions were highly imbalanced (Figure 3). Approximately 9.5% (1617/17,018) of the patients underwent KRT (1617 patients), resulting in an imbalance ratio of 9.5:1. Similarly, 18% (3063/17,018) of the patients died, corresponding to an imbalance ratio of 4.6:1.

Figure 3. Imbalanced outcome class distribution. (A) KRT; (B) in-hospital mortality. KRT: kidney replacement therapy.



Prediction of KRT

The predictive XGBoost model demonstrated high overall performance when considering accuracy (0.910) and AUROC (0.928; Table 1). However, due to class imbalance, the other metrics revealed some important observations that would otherwise be overlooked. Notably, the macro-F_1-score of 0.695 suggests relatively lower performance, specifically for the minority class (KRT=yes; Tables 1 and 2). This is mainly due to the low recall (0.372), indicating that the model struggles to correctly identify a large proportion of actual KRT cases. The precision for this class was 0.539, resulting in an F_1-score of 0.439 (Table 2).

Table 1. Global metrics at different cutoff thresholds for kidney replacement therapy[a].

Cutoff	Accuracy	AUROC[b]	Macro-F_1	Brier
50%	0.910 (0.906-0.914)	0.928 (0.923-0.933)	0.695 (0.682-0.708)	0.066 (0.063-0.069)
40%	0.907 (0.902-0.912)	0.930 (0.923-0.937)	0.711 (0.694-0.728)	0.062 (0.060-0.064)
30%	0.901 (0.896-0.906)	0.930 (0.923-0.937)	0.731 (0.719-0.743)	0.062 (0.060-0.064)
20%	0.890 (0.884-0.896)	0.930 (0.923-0.937)	0.742 (0.729-0.755)	0.062 (0.060-0.064)
10%	0.865 (0.861-0.869)	0.930 (0.923-0.937)	0.731 (0.723-0.739)	0.062 (0.060-0.064)

[a]Data are presented as mean (95% CI).
[b]AUROC: area under the receiver operating characteristic curve.

Table 2. Per-class metrics at different cutoff thresholds for kidney replacement therapy[a].

Cutoff	Precision		Recall		F_1	
	KRT[b]	No KRT	KRT	No KRT	KRT	No KRT
50%	0.539	0.936	0.372	0.966	0.439	0.951
	(0.508-0.570)	(0.934-0.938)	(0.345-0.399)	(0.961-0.971)	(0.415-0.463)	(0.949-0.953)
40%	0.511	0.942	0.442	0.956	0.474	0.949
	(0.481-0.541)	(0.938-0.946)	(0.406-0.478)	(0.953-0.959)	(0.442-0.506)	(0.946-0.952)
30%	0.480	0.953	0.563	0.936	0.518	0.945
	(0.460-0.500)	(0.950-0.956)	(0.536-0.590)	(0.933-0.939)	(0.495-0.541)	(0.942-0.948)
20%	0.449	0.967	0.701	0.910	0.547	0.937
	(0.430-0.468)	(0.963-0.971)	(0.668-0.734)	(0.905-0.915)	(0.525-0.569)	(0.934-0.940)
10%	0.400	0.981	0.837	0.868	0.542	0.921
	(0.390-0.410)	(0.978-0.984)	(0.817-0.857)	(0.864-0.872)	(0.529-0.555)	(0.918-0.924)

[a]Data are presented as mean (95% CI).
[b]KRT: kidney replacement therapy.

The confusion matrices for KRT prediction also highlight this trend, showing that lowering the threshold from 50% to 10% enhances sensitivity, evidenced by a 174.5% increase in true positive (from 51 to 140). However, this adjustment also leads to a 284.0% rise in false positive (from 50 to 192; Figure 4A).

Figure 4. (A) Confusion matrix at different cutoff thresholds to predict kidney replacement therapy. (B) Confusion matrix at different cutoff thresholds to predict in-hospital mortality.

A

Cutoff threshold = 50%

		True label: Yes	True label: No
Predicted label	Yes	51	50
	No	110	1491

Cutoff threshold = 40%

		True label: Yes	True label: No
Predicted label	Yes	72	64
	No	89	1477

Cutoff threshold = 30%

		True label: Yes	True label: No
Predicted label	Yes	98	94
	No	63	1447

Cutoff threshold = 20%

		True label: Yes	True label: No
Predicted label	Yes	127	131
	No	34	1410

Cutoff threshold = 10%

		True label: Yes	True label: No
Predicted label	Yes	140	192
	No	21	1349

B

Cutoff threshold = 50%

		True label: Yes	True label: No
Predicted label	Yes	202	88
	No	76	1336

Cutoff threshold = 40%

		True label: Yes	True label: No
Predicted label	Yes	217	107
	No	61	1317

Cutoff threshold = 30%

		True label: Yes	True label: No
Predicted label	Yes	232	128
	No	46	1296

Cutoff threshold = 20%

		True label: Yes	True label: No
Predicted label	Yes	251	149
	No	27	1275

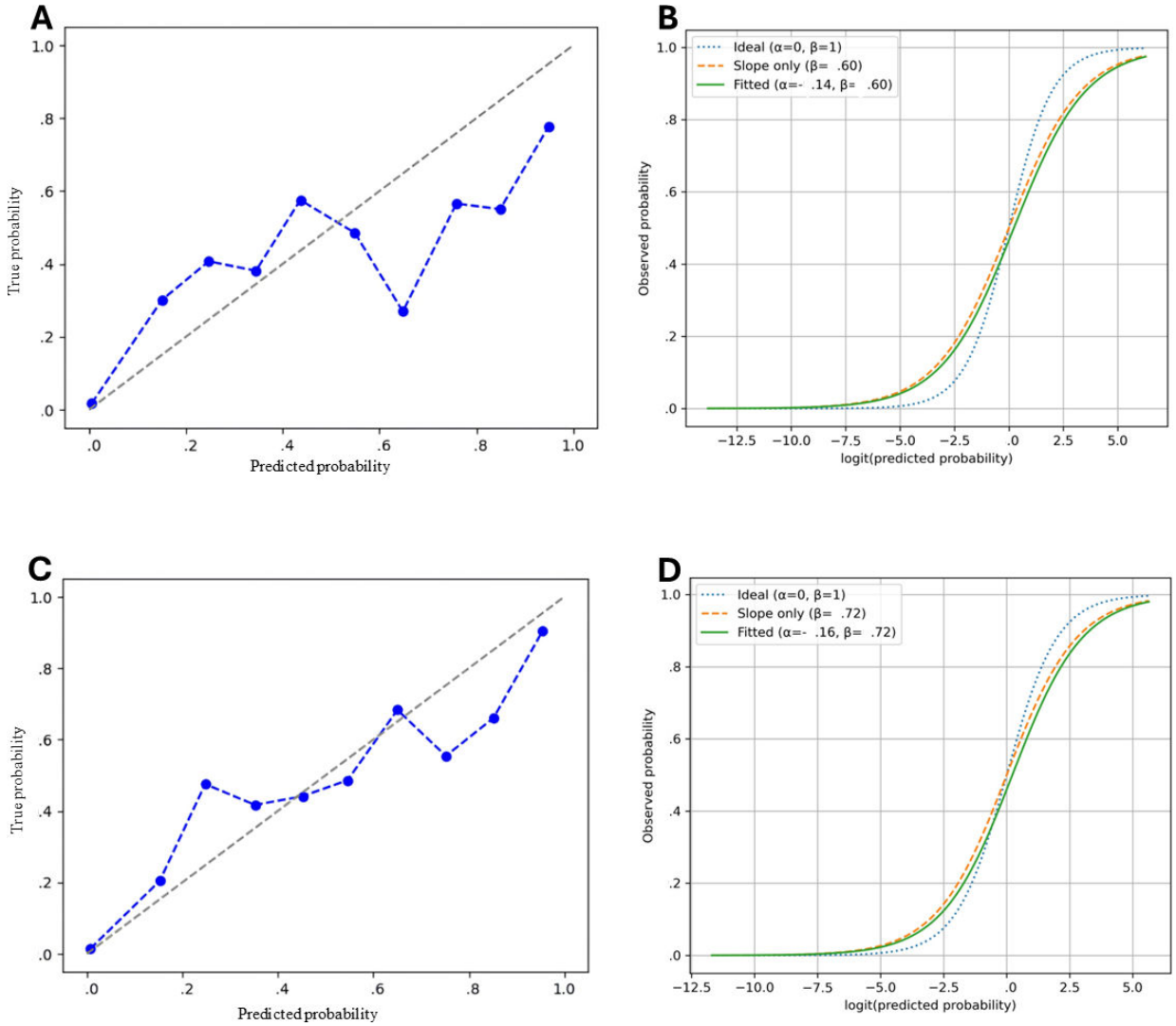
Cutoff threshold = 10%

		True label: Yes	True label: No
Predicted label	Yes	261	188
	No	17	1236

Changing the cutoff threshold affected the model's precision, recall, and F_1 values. In this context, considering the class of interest, which is KRT, and the default cutoff threshold of 50%, the precision was 0.539, recall was 0.372, and F_1-score was 0.439 (Table 2). Lowering the cutoff threshold to 20% resulted in a precision of 0.449, recall of 0.701, and an improved F_1-score of 0.547 (Table 2). The data presented in Multimedia Appendix 9 elucidate the trade-off between precision and recall, where an increase in precision usually implies a reduction in recall and vice versa.

The calibration plot shows systematic deviation from the diagonal, with predicted probabilities falling below the diagonal at higher values and above it at lower values, indicating overconfidence at the extremes (Figure 5A). This pattern is further supported by a calibration slope of 0.60 and an intercept of –0.14 (Figure 5B), indicating both overconfident predictions and a slight global overestimation of risk.

Figure 5. (A). Calibration curve for kidney replacement therapy. (B). Plot showing the calibration slope and intercept for the kidney replacement therapy task. (C). Calibration curve for death. (D). Plot showing the calibration slope and intercept for the death.

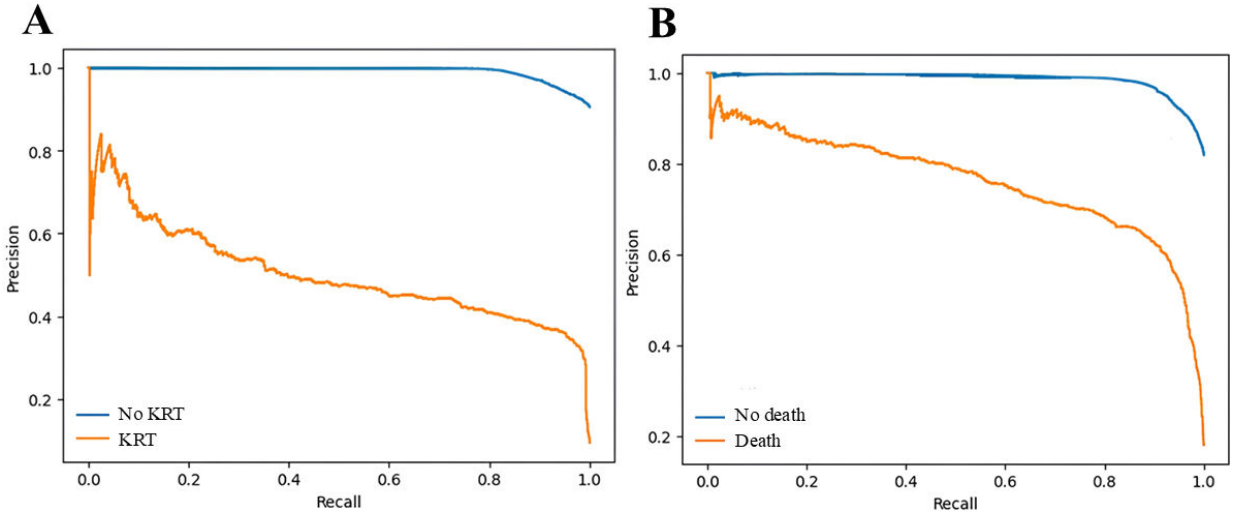


The precision-recall curve for each class (Figure 6A) showed that the non-KRT class achieved high precision and recall simultaneously, which is desirable, while the KRT class showed a performance closer to random. In other words, the model is relatively good at recalling non–KRT patients (high specificity), but it struggles to identify the ones who underwent KRT (low sensitivity).

It is observed that the curve for the class that did not undergo KRT remains close to the upper right corner, indicating that the model can achieve high precision and recall rates simultaneously. Conversely, the curve for the (interest) minority class (underwent KRT) approaches the diagonal, suggesting that the model is struggling to balance precision and recall, with performance close to random.

The decision curve analysis (Multimedia Appendix 10) indicates that the proposed model generates a positive net benefit for low to moderate decision thresholds, starting at approximately 0.10 and gradually decreasing to zero as the threshold increases to 0.5. Consequently, the model exhibits a practical benefit for decision thresholds ($P \leq .5$), indicating usefulness in scenarios that tolerate decisions based on relatively moderate predicted probabilities. In contrast, the strategy of treating all cases shows a positive net benefit only at very low thresholds, reaching zero around a threshold of 0.10 and becoming increasingly negative thereafter.

Figure 6. (A) Precision-recall curves for patients undergoing KRT and not undergoing KRT. (B) Precision-recall curves for death, and no death. KRT: kidney replacement therapy.



Prediction of In-Hospital Mortality

The predictive XGBoost model achieved high values of accuracy (0.900) and AUROC (0.945; Table 3). However, due to class imbalance, the macro-F_1-score of 0.830 indicates lower performance (Table 3). Specifically, for the minority class (deceased=yes), the precision is 0.725, the recall is 0.718, and the F_1-score is 0.721 (Table 4).

Table 3. Global metrics at different cutoff thresholds for death[a].

Cutoff	Accuracy	AUROC[b]	Macro-F_1	Brier
50%	0.900 (0.896-0.904)	0.945 (0.939-0.951)	0.830 (0.825-0.835)	0.072 (0.069-0.075)
40%	0.900 (0.895-0.905)	0.945 (0.939-0.951)	0.837 (0.830-0.844)	0.072 (0.069-0.075)
30%	0.896 (0.891-0.901)	0.945 (0.939-0.951)	0.837 (0.829-0.845)	0.072 (0.069-0.075)
20%	0.893 (0.887-0.899)	0.945 (0.939-0.951)	0.840 (0.831-0.849)	0.072 (0.069-0.075)
10%	0.878 (0.870-0.886)	0.945 (0.939-0.951)	0.827 (0.815-0.839)	0.072 (0.069-0.075)

[a]Data are presented as mean (95% CI).
[b]AUROC: area under the receiver operating characteristic curve.

The confusion matrices for mortality prediction demonstrated a similar precision-recall trade-off (Figure 4B). Reducing the cutoff threshold from 50% to 10% enhanced recall, with true positives increasing by approximately 29.1% (from 202 to 261), but also resulted in a 113% increase in false positives (from 88 to 188).

As with KRT, the cutoff threshold affected the precision, recall, and F_1-scores. At a 50% cutoff threshold, the precision is 0.725, the recall is 0.718, and the F_1-score is 0.721 (Table 4). Lowering the threshold to 20% increased the recall (0.880) and improved the F_1-score (0.748), while precision decreased to 0.651 (Table 4 and Multimedia Appendix 4). Similar to KRT, the calibration plot (slope=0.72; intercept=–0.16) was not satisfactory, and the Brier score was low (0.072; Figure 5C and D and Table 3).

The precision-recall curves for each class (Figure 6B) followed a similar pattern to that observed for KRT, with the non–death class exhibiting higher precision and recall than the death class. Once again, the model performed well in identifying survivors (high specificity) but demonstrated limited ability to detect patients who died (low sensitivity).

It is observed that the curve for the non–death class remains close to the upper right corner, indicating that the model can achieve high precision and recall rates simultaneously. In contrast, the curve for the death class, which is the minority class and of greater interest, approaches the diagonal (random model).

Similar to KRT, the decision curve for death shows that the net benefit of the strategy of treating all cases is approximately 0.2, while the proposed model achieves a substantially higher net benefit, around 0.8 (Multimedia Appendix 11). This behavior demonstrates that indiscriminate intervention quickly becomes inadequate as the decision threshold increases, while the proposed model maintains its practical usefulness over a substantially wider range of thresholds.

Table 4. Per-class metrics at different cutoff thresholds for death[a].

Cutoff	Precision		Recall		F_1	
	Death	No death	Death	No death	Death	No death
50%	0.725	0.938	0.718	0.940	0.721	0.939
	(0.703-0.747)	(0.934-0.942)	(0.710-0.726)	(0.934-0.946)	(0.712-0.730)	(0.937-0.941)
40%	0.701	0.949	0.776	0.927	0.736	0.938
	(0.678-0.724)	(0.946-0.952)	(0.768-0.784)	(0.920-0.934)	(0.725-0.747)	(0.935-0.941)
30%	0.673	0.958	0.821	0.912	0.739	0.935
	(0.652-0.694)	(0.955-0.961)	(0.811-0.831)	(0.906-0.918)	(0.726-0.752)	(0.931-0.939)
20%	0.651	0.971	0.880	0.896	0.748	0.932
	(0.630-0.672)	(0.969-0.973)	(0.872-0.888)	(0.889-0.903)	(0.734-0.762)	(0.928-0.936)
10%	0.608	0.980	0.921	0.869	0.732	0.921
	(0.585-0.631)	(0.977-0.983)	(0.910-0.932)	(0.861-0.877)	(0.714-0.750)	(0.916-0.926)

[a]Data are presented as mean (95% CI).

The Influence of Class Imbalance on Prediction

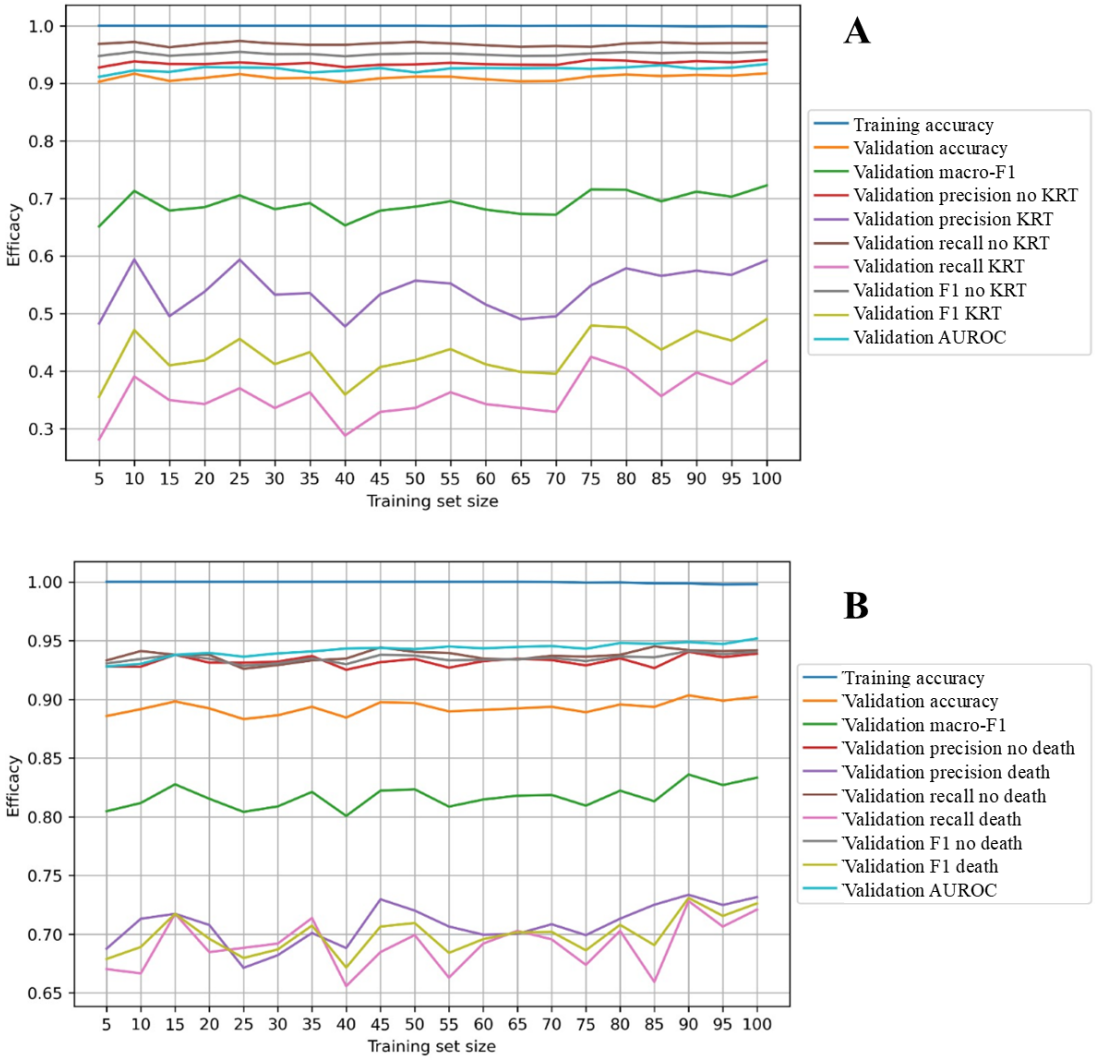
The model for KRT, which exhibited higher class imbalance, demonstrated superior performance for accuracy and AUROC when compared with the model for mortality. However, when examining the precision and recall for the minority class, the performance was suboptimal, and the KRT model exhibited lower performance than the mortality model. This pattern was also reflected in the macro-F_1-score, where the KRT model displayed a more significant drop in performance, further highlighting the impact of class imbalance.

It is important to highlight that KRT represents a distinct endpoint from mortality. Although both models used the same set of variables, the features and the importance of each feature vary depending on the endpoint (Multimedia Appendices 12 and 13). Therefore, discrepancies in the performance of the KRT and mortality models should not be solely attributed to differences in class balance.

Learning Curves Analysis

The learning curves show a plateau-like shape, with stable validation performance across all training set sizes for both outcomes (Figures 7A and 7B). This pattern suggests limited change in performance as the training set size increases, indicating that model performance does not substantially improve with additional data.

Figure 7. (A) Learning curves for different training set sizes for kidney replacement therapy. (B) Learning curves for different training set sizes for death. AUROC: area under the receiver operating characteristic curve; F_1: F_1-score; KRT: kidney replacement therapy.



Impact of Balancing Strategies on Prediction Results and on the Metrics

When evaluating the impact of class rebalancing strategies, we observe a clear and systematic discrepancy between AUROC and metrics that explicitly account for class-specific behavior. As shown in Multimedia Appendices 12 and 13, AUROC remains consistently high across all experimental conditions, exceeding 0.8 in every scenario, regardless of whether rebalancing is applied or whether minority-class performance substantially improves or deteriorates. This stability may give the misleading impression that rebalancing strategies have limited effect on model behavior.

However, a closer inspection using alternative metrics reveals a markedly different picture. In particular, TPRGap provides a direct measure of classifier bias induced by class imbalance, capturing disparities in true-positive rates between classes. Using this metric, several rebalancing techniques substantially reduce bias relative to the unbalanced baseline. For instance, in the death outcome, TPRGap decreases from 0.230 in the unbalanced setting to 0.043 when using Redundancy-Based Undersampling, indicating a pronounced reduction in class-dependent performance disparity. In contrast, AUROC changes only marginally in the same scenario, from 0.945 to 0.941, failing to reflect this improvement.

Similar patterns are observed across other class-aware metrics, particularly positive-class precision, recall, and F_1-score. These metrics exhibit significant sensitivity to rebalancing strategies, capturing both beneficial and harmful effects on minority-class performance. In several cases, rebalancing leads to meaningful gains in recall at the expense of precision, or vice versa, reflecting trade-offs that are critical in clinical decision-making contexts. Yet, AUROC remains largely unchanged, masking these trade-offs and providing little insight into how the classifier's behavior actually shifts.

An especially illustrative example arises in the death outcome under the Near Miss 2 undersampling strategy. In this case, the positive-class F_1-score drops sharply from 0.721 in the unbalanced model to 0.345, signaling a severe degradation in clinically relevant performance. Despite this substantial decline, AUROC remains comparatively high, decreasing from 0.945 to 0.809. This modest reduction does not adequately reflect the magnitude of the performance loss

experienced by the minority class, underscoring the disconnect between AUROC and clinically meaningful outcomes.

Discussion

Principal Findings

In recent years, the rapid expansion of ML applications in health care has led to an increasing number of predictive models being proposed for clinical use. However, many of these studies continue to rely primarily, or even exclusively, on AUROC to report model performance, even in highly imbalanced clinical scenarios. Our findings demonstrate the limitations of this practice, showing that high AUROC values may coexist with poor performance for clinically critical minority outcomes.

Although methodological literature has long acknowledged the limitations of AUROC and accuracy in imbalanced settings, clinical prediction studies frequently continue to emphasize these global metrics. By applying complementary class-specific measures and analyzing the learning curves in a large cohort of over 17,000 hospitalized patients with COVID-19, our study provides practical evidence of how metric selection directly influences clinical interpretation, particularly when outcomes are imbalanced.

For both KRT and in-hospital mortality, AUROC values suggested excellent discrimination. However, class-specific metrics revealed substantial deficiencies in identifying minority outcomes. The KRT model achieved an AUROC of 0.928. However, recall for the KRT class was only 0.372, meaning that the model failed to identify nearly two-thirds of patients who would require dialysis. In a real clinical scenario, such as a hospital without dialysis services, this could result in missed opportunities for early referrals, with serious consequences for patient care. This discrepancy was captured by the macro-F_1-score (0.695), which penalizes imbalanced performance across classes and thus offers a more clinically realistic summary of the model's performance than AUROC alone.

Calibration analysis further highlighted these limitations. Calibration slopes of 0.60 and 0.72, along with negative intercepts, indicate suboptimal calibration for both outcomes. Although Brier scores were low, this likely reflects strong discrimination combined with outcome imbalance, rather than well-calibrated probability estimates. These findings reinforce that no single metric adequately captures model performance in imbalanced clinical settings.

Precision-recall analysis further highlighted the limitations of AUROC-based evaluation by prioritizing performance on the minority class, which often represents the clinically most relevant outcome [57]. While decision curve analysis further illustrates that clinically meaningful usefulness may vary substantially across thresholds [61,62] and demonstrated net benefit across certain thresholds, it did not capture how prediction errors were distributed between classes. By examining class-specific precision and recall across varying thresholds, we directly linked model behavior to real-world clinical trade-offs between underdiagnosis and overdiagnosis. Given that class prevalence directly affects metrics such as precision and recall, particularly in imbalanced settings [11,61,63], relying on AUROC alone may obscure clinically relevant deficiencies in minority-class performance. These results underscore the importance of reporting complementary, class-aware metrics, as no single metric adequately captures model performance across different clinical contexts. Therefore, metric selection should be guided by the intended clinical application. In high-risk settings, maximizing recall may be preferable to avoid missing cases, even at the expense of increased false positives. In contrast, in resource-limited settings, higher precision may be prioritized to reduce unnecessary interventions. These trade-offs cannot be adequately captured by a single metric, reinforcing the need for a multidimensional evaluation approach.

Learning curve analysis provided an additional and complementary perspective on model performance. In general, learning curves may reflect either progressive improvement with increasing data or early convergence in relatively simple prediction tasks [60]. In our study, however, the combination of flat learning curves, persistently low minority-class performance, and stable AUROC values across increasing training set sizes suggests limited gains in performance as more data are added, highlighting that AUROC alone may not capture important aspects of model behavior in imbalanced settings.

This finding has relevant methodological considerations. Despite consistently high AUROC values (>0.92), the lack of substantial improvement in performance with increasing training data suggests that discrimination alone may not fully reflect how model performance evolves, particularly for identifying high-risk patients. In this context, AUROC may reflect stable ranking ability driven by dataset characteristics, such as class imbalance, rather than improvements in clinically relevant performance. Therefore, reliance on AUROC alone may lead to overestimation of model performance and potential underrecognition of high-risk patients.

Learning curves offer a complementary tool to assess how model performance changes as additional data are incorporated [60]. When performance remains relatively stable, caution is warranted in interpreting high discrimination metrics as sufficient evidence of model adequacy. Together, these findings reinforce that AUROC alone is insufficient to determine whether a model is suitable for clinical use, particularly in imbalanced scenarios.

International reporting frameworks such as PROBAST+AI and TRIPOD+AI emphasize comprehensive evaluation of predictive model performance and transparency in results presentation but remain focused on global performance measures of discrimination, calibration, and overall clinical usefulness [6,16]. Although these frameworks acknowledge the impacts of class imbalance on outcomes, they do not offer strategies for measuring the redistribution of errors across classes with varying thresholds, nor do they recommend

incorporating learning curve analysis into model evaluation [6,16].

Additionally, according to PROBAST+AI, applicability concerns were considered low, as the study population, predictors, and outcomes are consistent with real-world clinical settings. However, this should not be interpreted as evidence supporting clinical use of the models. Despite this apparent applicability, the models demonstrated important limitations in clinically relevant performance, including suboptimal calibration and limited sensitivity for the minority class. This apparent contradiction highlights a key finding of our study: even when models are developed using appropriate data and aligned with clinical contexts, reliance on conventional metrics such as AUROC may obscure critical weaknesses. Therefore, methodological soundness and contextual relevance alone are insufficient to ensure clinically meaningful performance, reinforcing the need for comprehensive, class-aware evaluation frameworks before considering any potential clinical implementation.

Our findings provide a methodological contribution that extends beyond these current standards. Specifically, we demonstrate that high AUROC values were maintained despite limited changes in performance across increasing training set sizes. This observation suggests that AUROC alone may not reflect whether a model has learned clinically meaningful patterns but rather may capture stable discrimination driven by dataset characteristics such as class imbalance, highlighting the need for more transparent and comprehensive reporting.

This has important implications: models may comply with current reporting standards while still underperforming in clinically relevant minority outcomes, which are often the most clinically relevant. Because learning curves are rarely reported, this limitation may go unrecognized in many published models. Incorporating learning curve analysis alongside class-specific metrics can therefore enhance transparency and provide a more robust assessment of model performance.

Resampling techniques, including over- and undersampling, resulted in only modest improvements in minority-class performance and did not resolve the fundamental limitations of AUROC-based evaluation [28,62,64,65]. Even when class distributions were artificially modified, AUROC remained largely insensitive to clinically meaningful changes in recall and precision. This reinforces that rebalancing alone cannot compensate for inappropriate performance metrics.

Our study contributes to literature by bridging theoretical concerns with practical, real-world application. By evaluating 2 predictive models in a large clinical cohort with varying degrees of outcome imbalance, we demonstrate how metric selection and threshold choice directly influence clinical interpretation and link how these shifts directly affect clinical decision-making. By evaluating per-class precision and recall across multiple cutoffs and visualizing these relationships through precision-recall curves, we make explicit the trade-offs inherent to real-world model deployment.

These trade-offs are particularly relevant in health care, where underdiagnosing high-risk patients (low recall) may lead to missed interventions, while overdiagnosis (low precision) can result in unnecessary procedures and resource strain [66]. In high-stakes settings, such as intensive care units or emergency triage, prioritizing recall may be appropriate, even at the cost of more false positives, to avoid missing patients at risk of deterioration. On the other hand, in resource-constrained environments, higher precision may be preferable. Together, these findings reinforce that no single metric is sufficient and that different clinical contexts require different operating points and different emphases on recall, precision, or their balance, which is effectively summarized by the macro-F_1.

Despite growing methodological awareness [4,6-9,50,64], most applied health care research still relies predominantly on this AUROC for model evaluation (Multimedia Appendices 14 and 15) [67]. For example, DynaMed, a widely used evidence-based clinical reference platform, currently lists 26 predictive models specifically developed for COVID-19—1 diagnostic and 25 prognostics, including outcomes such as severe disease progression, thrombosis, intensive care unit admission, KRT, and mortality (Multimedia Appendix 11) [67]. Notably, 88.5% (23/26) of these models primarily report AUROC as the main performance metric [67].

Additionally, some studies report multiple metrics without adequately contextualizing their relevance or the trade-offs involved [67-69]. Our findings highlight the importance of not only reporting multiple metrics but also interpreting them in relation to clinical context and outcome imbalance.

Therefore, metric selection and learning curve analysis may substantially influence the clinical interpretation of model performance. Choosing evaluation strategies that account for outcome imbalance and clinical priorities is essential to support more rigorous evaluation before potential clinical implementation of predictive models.

Limitations

The methodology focused on a single algorithm (XGBoost), although the observed patterns are not model-specific and reflect broader issues related to class imbalance and performance evaluation. External validation was not feasible due to data availability; however, this does not affect the central methodological contribution of the study, which concerns the interpretation of model performance rather than the generalizability of a specific model. Therefore, our findings should not be interpreted as supporting the clinical use of the models presented, given the lack of external validation and suboptimal calibration. Importantly, the aim of this study was not to develop the best-performing predictive model but to examine how evaluation strategies influence the interpretation of model performance in imbalanced clinical scenarios.

In prediction tasks with imbalanced outcomes, which are common in health care, reliance on accuracy and AUROC alone may obscure clinically important failures. Complementary metrics, including precision, recall, and macro-F_1, provide a more realistic assessment of model performance

and should be systematically reported. In addition, learning curve analysis offers insight into a model's learning dynamics and helps explore how model performance evolves as more training data are incorporated. Together, these approaches support a more comprehensive and clinically meaningful evaluation of predictive models, particularly in imbalanced settings, rather than their direct translation into clinical practice.

Acknowledgments

The authors would like to thank the hospitals which are part of this collaboration for supporting this project. They also thank all the clinical staff at those hospitals, who cared for the patients, and all undergraduate students who helped with data collection. The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: assisted linguistic editing and grammatical review. The GenAI tools used were Grammarly, Gemini 1.5, ChatGPT-5.2. The responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This study was supported in part by the Minas Gerais State Agency for Research and Development (Fundação de Amparo à Pesquisa do Estado de Minas Gerais–FAPEMIG) (grant APQ-01154-21), National Institute of Science and Technology for Health Technology Assessment (Instituto de Avaliação de Tecnologias em Saúde–IATS)/National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico–CNPq) (grant 408659/2024-6), and the Center for Innovation and Artificial Intelligence for Health (CI-IA Saúde), which is funded by the São Paulo State Research Support Foundation (FAPESP) (2020/09866-4), FAPEMIG (PPE-00030-21), and UNIMED Belo Horizonte. MSM was partially supported by CNPq (311742/2025-4).

Data Availability

The data generated or analyzed during this study are included in this paper and its Multimedia Appendices 1-18. The corresponding author is available to provide additional data regarding this manuscript upon reasonable request.

Authors' Contributions

Conception and design of the work: VGJV, MAG, MSM
Data collection: VGJV, BPP, CAP, HRV, MSM
Data curation: MSM
Data analysis and interpretation: VGJV, CMVA, JMA, GFN, LCDR, MAG, MSM
Drafting the paper: VGJV, CMVA, JMA, LCDR, MAG, MSM
Writing – review & editing: VGJV, CMVA, JMA, BPP, CAP, GFN, EB, HRV, KPF, LCDR, MAG, MSM
Project administration: MSM
Supervision: MSM
Reading and approving the final version of the manuscript: all authors

Conflicts of Interest

None declared.

Multimedia Appendix 1

Potential predictors for patients with COVID-19 undergoing kidney replacement therapy.
[DOCX File (Microsoft Word File), 2585 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Potential predictors for in-hospital mortality in patients with COVID-19.
[DOCX File (Microsoft Word File), 2590 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Final proportions in the training partition after application of each rebalancing strategy.
[DOCX File (Microsoft Word File), 3586 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Hyperparameters evaluated for optimization using extreme gradient boosting (XGBoost).
[DOCX File (Microsoft Word File), 2584 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Definitions and characteristics of the metrics frequently used to evaluate performance in predictive models, using the machine learning and the statistics terminology.

[DOCX File (Microsoft Word File), 2588 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Hypothetical confusion matrix and calculation of performance metrics for binary classification.
[DOCX File (Microsoft Word File), 2490 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Checklist for TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + Artificial Intelligence).
[DOCX File (Microsoft Word File), 2588 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Checklist for PROBAST+AI (Updated Quality, Risk of Bias, and Applicability Assessment Tool for Prediction Models Using Regression or Artificial Intelligence Methods).
[DOCX File (Microsoft Word File), 2097 KB-Multimedia Appendix 8]

Multimedia Appendix 9

Means of performance metrics at different cutoffs for kidney replacement therapy (KRT) and not undergoing KRT.
[PNG File (Portable Network Graphics File), 32 KB-Multimedia Appendix 9]

Multimedia Appendix 10

Decision curve for kidney replacement therapy.
[PNG File (Portable Network Graphics File), 255 KB-Multimedia Appendix 10]

Multimedia Appendix 11

Decision curve for death.
[PNG File (Portable Network Graphics File), 283 KB-Multimedia Appendix 11]

Multimedia Appendix 12

Global and per-class metrics for different rebalancing techniques for kidney replacement therapy.
[DOCX File (Microsoft Word File), 3130 KB-Multimedia Appendix 12]

Multimedia Appendix 13

Global and per-class metrics for different rebalancing techniques for death.
[DOCX File (Microsoft Word File), 2587 KB-Multimedia Appendix 13]

Multimedia Appendix 14

Outcomes and evaluation metrics of predictive scores for patients with COVID-19 based on the DynaMed summary.
[DOCX File (Microsoft Word File), 2586 KB-Multimedia Appendix 14]

Multimedia Appendix 15

Outcomes and evaluation metrics of predictive scores for cardiovascular disease based on the DynaMed summary.
[DOCX File (Microsoft Word File), 2583 KB-Multimedia Appendix 15]

Multimedia Appendix 16

Means of performance metrics at different cutoffs for death and no death.
[PNG File (Portable Network Graphics File), 28 KB-Multimedia Appendix 16]

Multimedia Appendix 17

Features' importance and contribution to the final predictive model of kidney replacement therapy.
[DOCX File (Microsoft Word File), 2580 KB-Multimedia Appendix 17]

Multimedia Appendix 18

Features' importance and contribution to the final predictive model of in-hospital mortality.
[DOCX File (Microsoft Word File), 2580 KB-Multimedia Appendix 18]

References

1. Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ. Jan 8, 2024;384:e074819. [doi: 10.1136/bmj-2023-074819] [Medline: 38191193]

2. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer Nature; 2019. URL: https://doi.org/10.1007/978-3-030-16399-0 [Accessed 2025-08-22]
3. van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. J Clin Epidemiol. Apr 2021;132:142-145. [doi: 10.1016/j.jclinepi.2021.01.009] [Medline: 33775387]
4. Adhikari S, Normand SL, Bloom J, Shahian D, Rose S. Revisiting performance metrics for prediction with rare outcomes. Stat Methods Med Res. Oct 2021;30(10):2352-2366. [doi: 10.1177/09622802211038754] [Medline: 34468239]
5. Cabot JH, Ross EG. Evaluating prediction model performance. Surgery. Sep 2023;174(3):723-726. [doi: 10.1016/j.surg.2023.05.023] [Medline: 37419761]
6. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. Apr 16, 2024;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
7. Cartus AR, Samuels EA, Cerdá M, Marshall BDL. Outcome class imbalance and rare events: an underappreciated complication for overdose risk prediction modeling. Addiction. Jun 2023;118(6):1167-1176. [doi: 10.1111/add.16133] [Medline: 36683137]
8. de Paiva BBM, Pereira PD, de Andrade CMV, et al. Potential and limitations of machine meta-learning (ensemble) methods for predicting COVID-19 mortality in a large inhospital Brazilian dataset. Sci Rep. Mar 1, 2023;13(1):3463. [doi: 10.1038/s41598-023-28579-z] [Medline: 36859446]
9. Liu S, Roemer F, Ge Y, et al. Comparison of evaluation metrics of deep learning for imbalanced imaging data in osteoarthritis studies. Osteoarthr Cartil. Sep 2023;31(9):1242-1248. [doi: 10.1016/j.joca.2023.05.006]
10. D'Agostino RB Sr, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. Feb 12, 2008;117(6):743-753. [doi: 10.1161/CIRCULATIONAHA.107.699579] [Medline: 18212285]
11. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 1st ed. John Wiley & Sons, Inc; URL: https://doi.org/10.1002/9781118548387 [Accessed 2025-08-22] [doi: 10.1002/9781118548387]
12. Iragorri N, Spackman E. Assessing the value of screening tools: reviewing the challenges and opportunities of cost-effectiveness analysis. Public Health Rev. 2018;39:17. [doi: 10.1186/s40985-018-0093-8] [Medline: 30009081]
13. Arnett DK, Blumenthal RS, Albert MA, et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association task force on clinical practice guidelines. Circulation. Sep 10, 2019;140(11):e596-e646. [doi: 10.1161/CIR.0000000000000678] [Medline: 30879355]
14. US Preventive Services Task Force, Krist AH, Davidson KW, et al. Behavioral counseling interventions to promote a healthy diet and physical activity for cardiovascular disease prevention in adults with cardiovascular risk factors: US Preventive Services Task Force recommendation statement. JAMA. Nov 24, 2020;324(20):2069-2075. [doi: 10.1001/jama.2020.21749] [Medline: 33231670]
15. Rufibach K. Use of Brier score to assess binary predictions. J Clin Epidemiol. Aug 2010;63(8):938-939. [doi: 10.1016/j.jclinepi.2009.11.009] [Medline: 20189763]
16. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. Mar 24, 2025;388:e082505. [doi: 10.1136/bmj-2024-082505] [Medline: 40127903]
17. Hageman S, Pennells L, Ojeda F, et al. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. Eur Heart J. Jul 1, 2021;42(25):2439-2454. [doi: 10.1093/eurheartj/ehab309]
18. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. Dec 16, 2019;17(1):230. [doi: 10.1186/s12916-019-1466-7] [Medline: 31842878]
19. Liu T, Krentz A, Lu L, Curcin V. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. Eur Heart J Digit Health. Jan 2025;6(1):7-22. [doi: 10.1093/ehjdh/ztae080] [Medline: 39846062]
20. Liu W, Laranjo L, Klimis H, et al. Machine-learning versus traditional approaches for atherosclerotic cardiovascular risk prognostication in primary prevention cohorts: a systematic review and meta-analysis. Eur Heart J Qual Care Clin Outcomes. Jun 21, 2023;9(4):310-322. [doi: 10.1093/ehjqcco/qcad017] [Medline: 36869800]
21. Andersen ES, Birk-Korch JB, Hansen RS, et al. Monitoring performance of clinical artificial intelligence in health care: a scoping review. JBI Evid Synth. Dec 1, 2024;22(12):2423-2446. [doi: 10.11124/JBIES-24-00042] [Medline: 39658865]
22. Oettl FC, Pareek A, Winkler PW, et al. A practical guide to the implementation of AI in orthopaedic research, part 6: how to evaluate the performance of AI research? J Exp Orthop. Jul 2024;11(3):e12039. [doi: 10.1002/jeo2.12039] [Medline: 38826500]

23. Hicks SA, Strümke I, Thambawita V, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep. Apr 8, 2022;12(1):5979. [doi: 10.1038/s41598-022-09954-8] [Medline: 35395867]
24. Megahed FM, Chen YJ, Megahed A, Ong Y, Altman N, Krzywinski M. The class imbalance problem. Nat Methods. Nov 2021;18(11):1270-1272. [doi: 10.1038/s41592-021-01302-4] [Medline: 34654918]
25. Lever J, Krzywinski M, Altman N. Classification evaluation. Nat Methods. Aug 2016;13(8):603-604. [doi: 10.1038/nmeth.3945]
26. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. Oct 29, 2019;17(1):195. [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
27. Kocak B, Klontzas ME, Stanzione A, et al. Evaluation metrics in medical imaging AI: fundamentals, pitfalls, misapplications, and recommendations. Eur J Radiol Artif Intell. Sep 2025;3:100030. [doi: 10.1016/j.ejrai.2025.100030]
28. Carriero A, Luijken K, de Hond A, Moons KGM, van Calster B, van Smeden M. The harms of class imbalance corrections for machine learning based prediction models: a simulation study. Stat Med. Feb 10, 2025;44(3-4):e10320. [doi: 10.1002/sim.10320] [Medline: 39865585]
29. Recommendations for national SARS-cov-2 testing strategies and diagnostic capacities: interim guidance, 25 June 2021. World Health Organization. 2021. URL: https://iris.who.int/handle/10665/342002 [Accessed 2025-08-22]
30. Marcolino MS, Ziegelmann PK, Souza-Silva MVR, et al. Clinical characteristics and outcomes of patients hospitalized with COVID-19 in Brazil: results from the Brazilian COVID-19 registry. Int J Infect Dis. Jun 2021;107:300-310. [doi: 10.1016/j.ijid.2021.01.019] [Medline: 33444752]
31. Yang L, Li J, Wei W, et al. Kidney health in the COVID-19 pandemic: an umbrella review of meta-analyses and systematic reviews. Front Public Health. 2022;10:963667. [doi: 10.3389/fpubh.2022.963667]
32. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. Apr 2009;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
33. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform. Jul 2019;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
34. Soriano Marcolino M, Minelli Figueira R, Pereira Afonso Dos Santos J, Silva Cardoso C, Luiz Ribeiro A, Alkmim MB. The experience of a sustainable large scale Brazilian telehealth network. Telemed J E Health. Nov 2016;22(11):899-908. [doi: 10.1089/tmj.2015.0234] [Medline: 27167901]
35. Bicalho MAC, Aliberti MJR, Delfino-Pereira P, et al. Clinical characteristics and outcomes of COVID-19 patients with preexisting dementia: a large multicenter propensity-matched Brazilian cohort study. BMC Geriatr. Jan 5, 2024;24(1):25. [doi: 10.1186/s12877-023-04494-w] [Medline: 38182982]
36. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2016:785-794; San Francisco, CA. [doi: 10.1145/2939672.2939785]
37. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. Inf Fusion. May 2022;81:84-90. [doi: 10.1016/j.inffus.2021.11.011]
38. Wang L, Wang X, Chen A, Jin X, Che H. Prediction of type 2 diabetes risk and its effect evaluation based on the XGBoost model. Healthcare (Basel). Jul 31, 2020;8(3):247. [doi: 10.3390/healthcare8030247] [Medline: 32751894]
39. Wilimitis D, Walsh CG. Practical considerations and applied examples of cross-validation for model development and evaluation in health care: tutorial. JMIR AI. Dec 18, 2023;2:e49023. [doi: 10.2196/49023] [Medline: 38875530]
40. Bradshaw TJ, Huemann Z, Hu J, Rahmim A. A guide to cross-validation for artificial intelligence in medical imaging. Radiol Artif Intell. Jul 2023;5(4):e220232. [doi: 10.1148/ryai.220232] [Medline: 37529208]
41. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Presented at: 14th International Joint Conference on Artificial Intelligence (IJCAI '95); Aug 20-25, 1995:1137-1145; Montreal, Canada. [doi: 10.5555/1643031.1643047]
42. Adin A, Krainski ET, Lenzi A, Liu Z, Martínez-Minaya J, Rue H. Automatic cross-validation in structured models: is it time to leave out leave-one-out? Spat Stat. Aug 2024;62:100843. [doi: 10.1016/j.spasta.2024.100843]
43. Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. arXiv. Preprint posted online on Sep 21, 2016. [doi: 10.48550/arXiv.1609.06570]
44. Garcia EA, He H, Bai Y, Li S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. Presented at: 2008 IEEE International Joint Conference on Neural Networks (IJCNN 2008); Jun 1-8, 2008:1322-1328; Hong Kong, China. [doi: 10.1109/IJCNN.2008.4633969]
45. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002;16:321-357. [doi: 10.1613/jair.953]

46. Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Huang DS, Zhang XP, Huang GB, editors. Advances in Intelligent Computing. Springer Nature; 2005:878-887. URL: https://doi.org/10.1007/11538059_91 [Accessed 2026-01-28] [doi: 10.1007/11538059_91]
47. Nguyen HM, Cooper EW, Kamei K. Borderline over-sampling for imbalanced data classification. Int J Knowledge Eng Soft Data Paradigms. 2011;3(1):4. [doi: 10.1504/IJKESDP.2011.039875]
48. Douzas G, Bacao F, Last F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Inf Sci (Ny). Oct 2018;465:1-20. [doi: 10.1016/j.ins.2018.06.056]
49. More A. Survey of resampling techniques for improving classification performance in unbalanced datasets. arXiv. Preprint posted online on Aug 22, 2016. [doi: 10.48550/arXiv.1608.06048]
50. Murphy KP. Machine Learning: A Probabilistic Perspective. MIT Press; 2012. ISBN: 0262018020
51. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol. 2020:37-63. URL: https://bioinfopublication.org/files/articles/2_1_1_JMLT.pdf [Accessed 2026-06-17] [doi: 10.9735/2229-3981]
52. Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics. Jan 1, 2012;28(1):112-118. [doi: 10.1093/bioinformatics/btr597] [Medline: 22039212]
53. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. Jul 2009;45(4):427-437. [doi: 10.1016/j.ipm.2009.03.002]
54. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. Presented at: Proceedings of the 23rd international conference on Machine learning - ICML '06; Jun 25-29, 2006:233-240; Pittsburgh, PA. [doi: 10.1145/1143844.1143874]
55. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-574. [doi: 10.1177/0272989X06295361] [Medline: 17099194]
56. Lipton ZC, Elkan C, Naryanaswamy B. Optimal thresholding of classifiers to maximize F1 measure. Mach Learn Knowl Discov Databases. 2014;8725:225-239. [doi: 10.1007/978-3-662-44851-9_15] [Medline: 26023687]
57. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology (Sunnyvale). Jan 2010;21(1):128-138. [doi: 10.1097/EDE.0b013e3181c30fb2] [Medline: 20010215]
58. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. J Am Med Inform Assoc. Apr 1, 2020;27(4):621-633. [doi: 10.1093/jamia/ocz228] [Medline: 32106284]
59. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. JAMA. Oct 10, 2017;318(14):1377-1384. [doi: 10.1001/jama.2017.12126] [Medline: 29049590]
60. Viering T, Loog M. The shape of learning curves: a review. IEEE Trans Pattern Anal Mach Intell. Jun 2023;45(6):7799-7819. [doi: 10.1109/TPAMI.2022.3220744] [Medline: 36350870]
61. Simon GJ, Aliferis C, editors. Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls. Springer International Publishing; 2024. URL: https://doi.org/10.1007/978-3-031-39355-6 [Accessed 2026-04-10] [doi: 10.1007/978-3-031-39355-6]
62. Ke JXC, DhakshinaMurthy A, George RB, Branco P. The effect of resampling techniques on the performances of machine learning clinical risk prediction models in the setting of severe class imbalance: development and internal validation in a retrospective cohort. Discov Artif Intell. 2024;4(1):91. [doi: 10.1007/s44163-024-00199-0] [Medline: 39624046]
63. Dinov ID. Data Science and Predictive Analytics: Biomedical and Health Applications Using R. Springer; 2018. URL: https://link.springer.com/book/10.1007/978-3-031-17483-4 [Accessed 2026-04-10]
64. Welvaars K, Oosterhoff JHF, van den Bekerom MPJ, Doornberg JN, van Haarst EP, OLVG Urology Consortium, and the Machine Learning Consortium. Implications of resampling data to address the class imbalance problem (IRCIP): an evaluation of impact on performance between classification algorithms in medical data. JAMIA Open. Jul 2023;6(2):ooad033. [doi: 10.1093/jamiaopen/ooad033] [Medline: 37266187]
65. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inform Assoc. Aug 16, 2022;29(9):1525-1534. [doi: 10.1093/jamia/ocac093] [Medline: 35686364]
66. Monaghan TF, Rahman SN, Agudelo CW, et al. Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value. Medicina (Kaunas). May 16, 2021;57(5):503. [doi: 10.3390/medicina57050503] [Medline: 34065637]
67. Clinical criteria. DynaMed. Aug 22, 2025. URL: https://www.dynamed.com/calculators/#cc-idx [Accessed 2026-05-14]

68. Riley RD, Pate A, Dhiman P, Archer L, Martin GP, Collins GS. Clinical prediction models and the multiverse of madness. BMC Med. Dec 18, 2023;21(1):502. [doi: 10.1186/s12916-023-03212-y] [Medline: 38110939]
69. Bozkurt C, Aşuroğlu T. Mortality prediction of various cancer patients via relevant feature analysis and machine learning. SN Comput Sci. 2023;4(3):264. [doi: 10.1007/s42979-023-01720-5]

Abbreviations

AUROC: area under the receiver operating characteristic curve
e2sc-us: Effective, Efficient, and Scalable Confidence-Based-UnderSampling
KRT: kidney replacement therapy
ML: machine learning
PROBAST+AI: Updated Quality, Risk of Bias, and Applicability Assessment Tool for Prediction Models Using Regression or Artificial Intelligence Methods
REDCap: Research Electronic Data Capture
SMOTE: Synthetic Minority Over-Sampling Technique
TRIPOD+AI: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis + Artificial Intelligence
XGBoost: extreme gradient boosting

Edited by Javad Sarvestan; peer-reviewed by Milan Toma, Nadia Abd-Alsabour; submitted 01.Nov.2025; final revised version received 20.Apr.2026; accepted 22.Apr.2026; published 24.Jun.2026

Please cite as:
Ventura VdGJ, Andrade CMV, Almeida JM, Pessoa BP, Polanczyk CA, Nascimento GFdo, Boersma E, Vianna HR, Farah KdP, Rocha LCDda, Gonçalves MA, Marcolino MS
Beyond Area Under the Receiver Operating Characteristic Curve: Evaluating Predictive Performance Metrics Under Class Imbalance in Real-World Clinical Data
JMIR Form Res 2026;10:e86379
URL: https://formative.jmir.org/2026/1/e86379
doi: 10.2196/86379